Hybrid Alchemical Free Energy/Machine-Learning Methodology for the Computation of Hydration Free Energies

A Hybrid Alchemical Free Energy/Machine Learning Methodology for the Computation of Hydration Free Energies

Jenke Scheen,[†] Wilson Wu,[†] Antonia S. J. S. Mey,[†] Paolo Tosco,[‡] Mark Mackey,[‡] and Julien Michel[*,†]

[†]*EaStCHEM School of Chemistry, University of Edinburgh, David Brewster Road, Edinburgh EH9 3FJ, United Kingdom*

[‡]*Cresset Group, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, United Kingdom*

E-mail: mail@julienmichel.net

Abstract

A methodology that combines alchemical free energy calculations (FEP) with machine learning (ML) has been developed to compute accurate absolute hydration free energies. The hybrid FEP/ML methodology was trained on a subset of the FreeSolv database, and retrospectively shown to outperform most submissions from the SAMPL4 competition. Compared to pure machine-learning approaches, FEP/ML yields more precise estimates of free energies of hydration, and requires a fraction of the training set size to outperform standalone FEP calculations. The ML-derived correction terms are further shown to be transferable to a range of related FEP simulation protocols. The approach may be used to inexpensively improve the accuracy of FEP calculations, and to flag molecules which will benefit the most from bespoke forcefield parameterization efforts.

Introduction

Alchemical free energy calculations (or Free Energy Perturbation -FEP-) are increasingly used in academia and industry to support ligand optimisation problems in the early stage of drug discovery.[1–4] The domain of applicability of current alchemical methodologies has to date mainly been restricted to hit-to-lead and lead optimisation scenarios owing to limitations in computing cost, conformational sampling, and the accuracy of the potential energy functions used to compute protein-ligand energetics.[5–10] There is continued interest in the development of more accurate potential energy functions to benchmark FEP workflows on diverse well-curated protein-ligand datasets,[11–14] and for applications to blinded challenges or methodological studies.[15–20]

The calculation of hydration free energies has historically been an important stepping stone towards more accurate forcefields for protein-ligand binding free energy calculations.[21–23] Blinded competitions such as SAMPL have also focused on hydration free energy calculations.[24] Forcefield parameterization is a painstaking challenge that requires meticulous and laborious efforts to yield steady gains in accuracy. Recent parameterization efforts from the Open Force Field, AMBER, CHARMM communities have involved multiple groups.[25–27] Recent work has sought to simplify the parameterization process by direct chemical perception of hierarchical parameter types.[28] Nevertheless it can be difficult to identify what modifications to introduce to improve the accuracy of parameter sets. Ultimately fundamental limits in accuracy cannot be overcome due to an incomplete description of the physics of the process, for instance due to use of fixed-charge forcefields that neglect polarisation effects.[29] Notably this realisation has prompted the development of post-processing methodologies based on quantum mechanical (QM) calculations to introduce correction terms for hydration and binding free energies computed by FEP methods using a classical force field.[25–27,30,31] Data-driven machine-learning (ML) methods have witnessed a resurgence of interest in drug discovery in recent years. Impressive advances have been made in the area of machine learning of quantum chemical calculations,[32,33] virtual screening,[34,35] and free energies of

hydration.[36–39] Efforts such as DeepChem[40] and MoleculeNet[41] have popularised the use of ML methods for molecular property predictions. Recent efforts have made use of 3D convolutional neural networks or other graph convolutional neural networks to predict binding affinities from the spatial structure of protein-ligand systems.[42,43] While impressive results have been demonstrated, the performance of ML methods is limited by the requirements of often substantial training sets, and a rapid decrease in accuracy when applying the models to molecules that are dissimilar to those that were included in the training set.

In previous work undertaken by our group as part of the SAMPL6 competition[20] we observed that empirically correcting FEP-derived host-guest binding free energies by a linear regression model calibrated on preceding SAMPL5 submissions,[44] led to significant decrease in mean unsigned error (MUE) of the predicted binding affinities. The present study extends this approach with machine-learning regression models that act as empirical correction terms to the FEP results. That is, the ML models are trained to predict the *mistake* compared to experimental values in Gibbs free energy that alchemical calculations make, referred to from here on as the ΔG_{offset}.

For any given alchemical prediction ΔG_{FEP} and associated experimental free energy ΔG_{EXP}, ΔG_{offset} is defined as the difference between the two; it also constitutes the training label for the given perturbation. This method relies on the assumption that given a training set of sufficient size, an empirical model trained on this set will be able to estimate accurately ΔG_{offset} values for a new set of alchemical predictions, thereby compensating for systematic errors in the underlying alchemical methodology.

As a proof-of-concept, we explore absolute alchemical calculations of hydration free energies performed with GROMACS.[45] Our results show that the proposed hybrid FEP/ML methodology leads to significant improvements in the accuracy of calculated hydration free energies, whilst only requiring modest training sets compared to a pure machine learning approach.

Theory & methods

FEP/ML model generation

The present methodology describes a regression model that fits the *mistake* that an alchemical calculation makes for a given molecule A, where the mistake is defined by equation 1:

$$\Delta G_{offset}(A) = \Delta G_{EXP}(A) - \Delta G_{FEP}(A), \tag{1}$$

where $\Delta \mathrm{G}_{FEP}(A)$ is the hydration free energy of molecule A calculated by the alchemical method, and $\Delta \mathrm{G}_{EXP}(A)$ is the experimentally determined hydration free energy for the same molecule. For a given training set with defined descriptors, machine-learning models were used to fit the training domain using five-fold cross-validation over 10 replicates, resulting in a total population N_{pop} of 50 trained models (see methods section below). All individual models in N_{pop} are regression models predicting their own $\Delta \hat{G}_{offset}$ value. We define our offset estimator as the arithmetic mean of these offset values, and use the standard deviation of the mean as a measure of the precision of the calculated offset. Thus we define a corrected hydration free energy as:

$$\Delta G_{FEP/ML}(A) = \Delta G_{FEP}(A) + \langle \Delta \hat{G}_{offset}(A) \rangle_{N_{pop}}. \tag{2}$$

and the precision of the $\Delta \mathrm{G}_{FEP/ML}(A)$ estimate is determined by propagating statistical errors of the alchemical and ML terms.

Dataset acquisition

Version 0.52 of the FreeSolv database[46] was downloaded from `https://github.com/MobleyLab/FreeSolv`. This version contains 642 small neutral molecules. Aside from experimentally-determined values, the database contains absolute free energies of hydration computed from alchemical simulations using GROMACS.[45] A detailed description of the particular FEP

methodology used can be found in Ramos Matos et al. [47] FreeSolv calculations were performed using the GAFF[48] force field, AM1-BCC[49] partial charges and the TIP3P water model.[50,51]

A dataset split was performed by excluding the FreeSolvSAMPL4 set which contains all the compounds (n=47) that were used in the SAMPL4 blinded competition (and had been subsequently appended to the FreeSolv database after this challenge).[52] Compounds belonging to this set were extracted by filtering for the keyword 'SAMPL4_Guthrie' in the experimental reference column of the database's overview textfile. Six molecules (mobley_6309289, mobley_3395921, mobley_6739648, mobley_2607611, mobley_637522 and mobley_172879) were added manually to the test set because even though they were present in the SAMPL4 challenge they were not tagged with this keyword in v0.52 of the FreeSolv database. This resulted in a training set of 595 molecules. From here on only the training set will be described, but all treatment of data can be considered equal between the training and test set unless otherwise indicated. All data-handling was done in Python 3.7.4.

Feature generation & pre-processing

Features (descriptors) were generated for all compounds present in FreeSolv. The ML models in this study were generated using RDKit 2019.03.4.0.[53] Molecules were loaded using the provided SDF files, and featurized using the following classes on standard settings unless indicated otherwise:

- *APFP*: Atom-pair fingerprints were generated using `rdkit.Chem.rdMolDescriptors.GetHashedAtomPairFingerprint()`; length was set to 256.

- *ECFP*: Extended-connectivity fingerprints were generated using `rdkit.Chem.AllChem.GetMorganFingerprintAsBitVect()`; length was set to 1024. In order to generate fingerprints with diameters ECFP2/4/6/8, the radius was set to 1, 2, 3 and 4, respectively.

- *TOPOL*: Topological fingerprints were generated using `rdkit.Chem.RDKFingerprint()`; length was set to 1024.

- *MolProps*: Molecular properties were generated using the Mordred python API[54] with inclusion of 3D properties. Although the total number of descriptors that this API generates is 1825, non-numeric columns were excluded resulting in 1113 properties that constitute the features per compound. This particular molecular properties generator was chosen owing to the large number of molecular properties readily computed via its API.

- *X-NOISE*: Noise 'fingerprints' were generated using `NumPy.random.randint()`; length was set to 100 and random integers ranged between 0-100.

Additionally, all fingerprints were appended individually to MolProps features (resulting in for instance a feature set called 'MolPropsAPFP' which was obtained by appending 'APFP' to 'MolProps') resulting in fingerprints with a length of the sum of both feature sets (in the case of MolPropsAPFP, 1113 + 256 = 1369). Every feature set was subsequently Z-normalized to zero mean and `sklearn.decomposition.PCA` was used to reduce dimensionality using a principal component analysis, and retaining principal components contributing up to 95% of the variance.

After data pre-processing, the corresponding label (ΔG_{offset}, see Eq. 1) was appended to each data point in order to build the final training set (named 'FEP/ML'). Additionally, a second training set (named 'ML') was generated by using as labels (output variables) the experimentally-determined ΔG_{exp} value for each data point.

A 5-fold cross-validation approach was chosen to reduce the risks of overfitting the training set. The training set was thus randomly split into five equally-sized folds (of sizes 595/5=119). Training was repeated five times, rotating the folds so that each fold acted as the validation set once for the other four training set folds. Additionally, training was performed with 10 replicates per feature set, resulting in a total of 50 trained models per

feature set-ML model combination.

Machine-learning models

Scikit-Learn 0.11.1[55] was used to generate all ML models. The models were generated on a machine running Ubuntu 18.04.3 LTS containing 10 Inter i9-7900X CPU cores. For Support vector machines (SVMs), random forests (RFs), deep neural networks (DNNs) and multiple linear regressions (MLRs), the classes `sklearn.svm.SVR`, `sklearn.ensemble.RandomForestRegressor`, `sklearn.neural_network.MLPRegressor` and `sklearn.linear_model.LinearRegression` were used on standard settings except for DNN which used `max_iter=5000`.

In order to choose optimal hyperparameter configurations for each ML model, a Bayesian hyperparameter optimization routine was adopted using SciKit-Optimize 0.5.2 (SKOPT),[56] which makes use of an *expected improvement* acquisition function to search hyperparameter space more efficiently than a random or grid search. The number of steps (*calls* in SKOPT nomenclature) was set to 50 because convergence was observed before this point in most cases. After training a call, the cost function (mean absolute error of predicting on the validation set) across folds is returned to the SKOPT decorator which in turn chooses a new hyperparameter configuration for the next call using its acquisition function to attempt to further decrease the model's cost function. A more detailed description of the algorithm can be found in the online SKOPT documentation. Note that this means that for any ML model, each of the 10 replicates had its own configuration of hyperparameters, but within each replicate all five folds would have the same hyperparameter configuration. The complete hyperparameter space is described in table S1. Approximate runtimes for the complete training protocols were, for SVM, MLR, DNN and RF, 10h, 25h, 104h and 134h, respectively.

The code to reproduce all key results and figures presented in this manuscript is available at `https://github.com/michellab/hybrid_FEP-ML`.

Results & discussion

Protocol optimization on training set

For all the ML models derived in this study it was observed that hyperparameters played an important role in model validation accuracy. This is likely due to the relatively small size of the training set (595 datapoints). Thus a hyperparameter optimization algorithm was adopted in which hyperparameters were tuned with the help of Bayesian optimization based on Gaussian process regression (see table S1). This algorithm searches through hyperparameter space by wrapping around noisy, expensive ML functions; after 50 calls (configuration attempts), the hyperparameter configuration returning the lowest validation error is saved together with the corresponding trained model. For (SVM, RF and DNN models convergence was observed from around 30 calls. MLR in this case does not have any hyperparameters to tune which means that in every SKOPT call the same model is trained which results in an equal validation error along calls.

Based on the training protocol it can be observed that random forests (RF) and multiple linear regressions (MLR) do not fit the training set as well as support vector machines (SVM) and deep neural networks (DNN) protocols (Figure S1; SVM and DNN converge to validation MUE as low as $\sim$0.55 kcal·mol^{-1} whereas MLR and RF converge to $\sim$0.75 kcal·mol^{-1}.). For MLR this is to be expected because of the relative simplicity of the model. Although the RF algorithm is more complex, it is primarily designed for classification problems rather than regression problems due to its dependence on decision trees which may explain its underfitting. The algorithm is included as a control in the current study.

A range of different feature sets was used to identify efficient encodings for describing ΔG_{offset}. A general trend in feature set performance can be observed across ML models. MolProps and combinatorial feature sets (fingerprints appended to MolProps) fit the training set better than standalone fingerprints (APFP, TOPOL and ECFP6), and X-NOISE performs worst as expected since this feature set is generated from random data.

Because standalone MolProps generally outperform standalone fingerprints, it is likely that the combined feature sets benefit mainly from the more predictive MolProps component. The observation that MolProps appears to outperform other feature sets suggests some of the descriptors included in MolProps correlate well with free energies of hydration. This is reinforced by our observation that the MolProps feature set outperforms generally other feature sets when predicting ΔG of hydration directly in our pure ML models (see figure S1).

Although the Extended-connectivity fingerprint (ECFP)[57] is used extensively in QSAR regression problems, our training protocol suggests underfitting of the training set for this feature type. This is likely because the used diameter of six bonds is too large to accurately discriminate between the relatively small compounds in the FreeSolv database (see figure S4; typical MW is 100 Da); testing with smaller diameters suggests an increase in fitting ability, however these models still underperform with respect to other feature types (see figure S3; ECFP8,6,4,2 converge to 0.83, 0.82, 0.78 and 0.67 kcal·mol^{-1}, respectively).

Hybrid FEP/ML models outperform standalone FEP and ML models in SAMPL4

The trained models were used to predict on the Freesolv-SAMPL4 test set. Because low errors in training validation do not necessarily translate into low errors in testing validation, all trained models were tested (see figure S2 and table S2). Top-performing models per ML model (see figure 1) were based primarily on the MolProps feature set for SVM, RF and MLR, but not for DNN. It is likely that the latter suffers from a degree of overfitting causing individual models to differ widely in predicted offset values. This is apparent in the much larger uncertainties in dataset metrics for DNN. Nevertheless the accuracy of the predictions obtained by averaging over the 50 DNN models is competitive. Overall SVM appeared to give more consistently accurate and precise estimates of ΔG_{offset} values.

One compound in the test set (mobley_4587267, (2R,3R,4R,5R)-hexane-1,2,3,4,5,6-hexol,

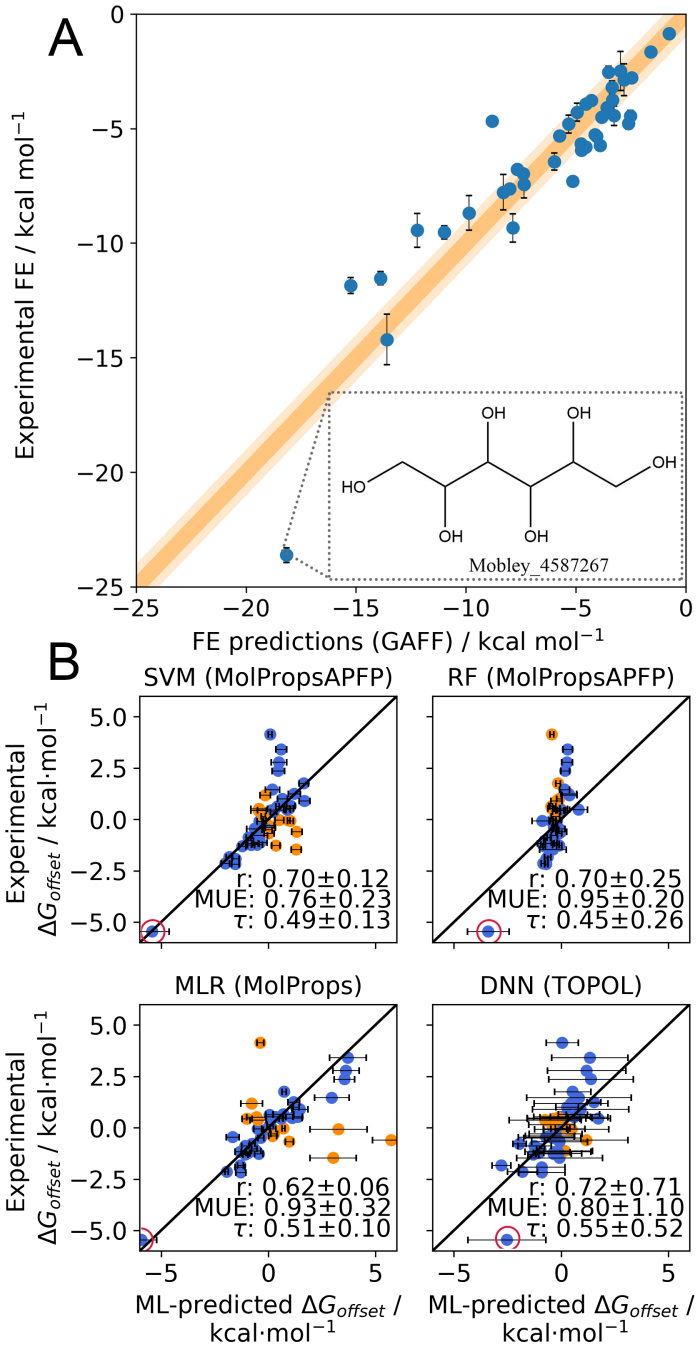

Figure 1: *Overview of prediction results on the SAMPL4-Freesolv test set.* **A:** FEP-predicted versus experimentally-determined free energies of hydration in kcal·mol^{-1}. The orange and light-orange areas are confidence regions for 1 and 2 kcal·mol^{-1}, respectively. Statistical uncertainties as supplied by the authors are shown as errorbars. **B:** Scatter plots of top-performing ML models predicting ΔG_{offset} for the FreeSolvSAMPL4 set with respective statistical intervals. Corrections with correct directionality (i.e. when $\langle \Delta \hat{G}_{offset} \rangle_{N_{pop}}$ and ΔG_{offset} values are both positive or both negative) are shown in blue; Corrections with incorrect directionality are shown in orange. The error bars on x-axis values denote the standard error of the mean offset value from ensembles of 50 ML models. Black diagonal lines show the $x = y$ diagonals.

referred to as mannitol from hereon) stands out with a free energy of hydration significantly more negative than other compounds in the test set ($\sim$ -24 kcal·mol^{-1}). This compound has a large associated ΔG_{offset} value of $\sim$-5 kcal·mol^{-1} (figures 1A and 1B, resp.). SVM and MLR models appear to correct this outlier better than RF and DNN models do, and it is likely that this outlier correction skews the statistical performances of the four models to a degree (see table S3 for model performances excluding the outlier); indeed, when plugging in the correction terms (figure 2), FEP/ML FE predictions for mannitol appear to be close to experimental hydration free energy measures, especially for SVM and MLR models.

The top-performing FEP/ML model (SVM; MolPropsAPFP, Figure 2A) outperformed standalone FEP in Pearson r, MUE and RMSE statistics (0.96, 0.76 kcal·mol^{-1} and 1.28 kcal·mol^{-1} respectively for FEP/ML; 0.92, 1.16 kcal·mol^{-1} and 2.49 kcal·mol^{-1} respectively for standalone FEP) and had higher ranking statistics (Spearman ρ and Kendall τ) than standalone FEP (see table S2; 0.96, 0.86 respectively for FEP/ML; 0.87, 0.72 respectively for standalone FEP). The top-performing FEP/ML DNN model achieves similar accuracy, but introduces significant uncertainties compared to FEP (Figure 2B). This reflects the larger uncertainties of the DNN-derived offset values in comparison with other ML protocols (see Figure 1B). Even when offset predictions for a given model are of modest accuracy, plugging in the correction term results in a FEP/ML model free energy prediction that performs equally well than the standalone FEP component. It seems that instead of predicting increasingly random values, the worse ΔG_{offset} predictor models converge towards predicting the training set mean offset value (-0.32 kcal·mol^{-1}) for all compounds (see table S2 X-NOISE entries). This is significant because it implies that, given that a properly-trained model is used, the correction term can be applied confidently to FEP datasets with minimal risk of worsening the model performance. The exception to this observation is MLR, which appears to occasionally predict large ΔG_{offset} values greater than 10 kcal·mol^{-1}. This was confirmed by high training validation values in figure S1 and bottom-level FEP/ML entries in table S2.

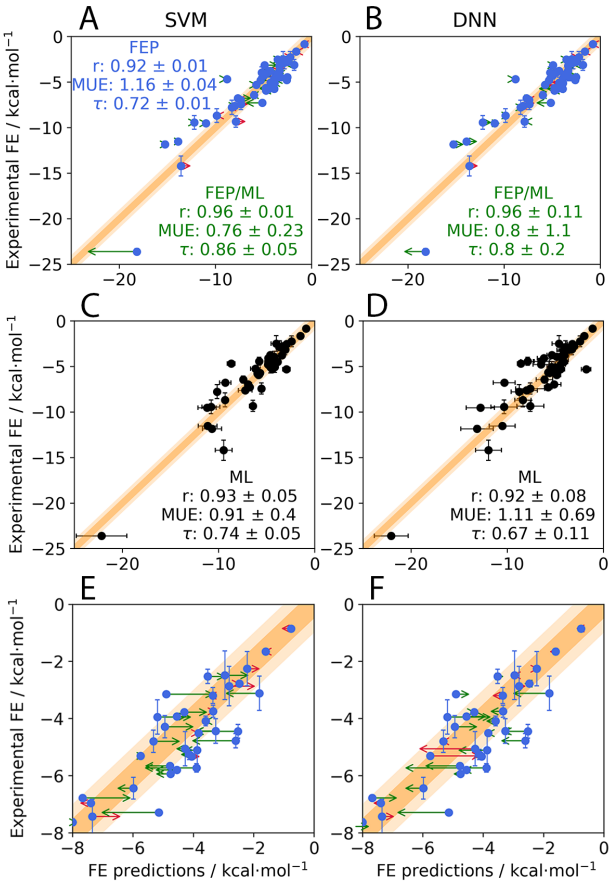

Figure 2: *Machine-learned correction terms applied to FEP predictions.* Results are shown for both support vector machine (left column and deep neural network (right column) ensembles. **A/B:** The FreeSolvSAMPL4 set FEP predictions (figure 1) with corrections as predicted by ML models shown with arrows. Green/red arrows depict corrections that improve/worsen agreement with experiment. Statistics for standalone FEP (blue) and hybrid FEP/ML (green) are shown. **C/D:** pure machine-learning (ML) models directly predicting ΔG of hydration with statistics in black text. **E/F:** contains the same data as A/B, but with a smaller range on both axes. Model uncertainties are shown as error bars. For all statistics the uncertainties are shown with a plus-minus sign.

The top-performing ML model (SVM; MolProps, Figure 2C) achieves accuracy similar to FEP, but with larger uncertainties. This trend worsens for the top-performing DNN model (Figure 2D). As noted before, mannitol contributes substantially to model performance: a second table with statistical performances excluding mannitol can be found in table S3. Indeed, excluding this compound slightly diminishes the gain in performance when comparing FEP/ML models to standalone FEP, although ranking statistics seem to benefit equally well from correction compared to when mannitol is included. This suggests that the small corrections (figure 2E and F) introduce primarily a correct reordering of compound ΔG values.

The top-performing FEP/ML (SVM; MolPropsAPFP)and ML (SVM; MolProps) models were introduced in the SAMPL4 challenge retrospectively (figure 3) to correct the results of SAMPL4 submission 004 that featured a FEP protocol most similar to the one used to generate calculated FEP values in FreeSolv. In line with the results obtained on the FreeSolvSAMPL4 test set, FEP/ML SVM models trained with MolPropsAPFP outperformed standalone FEP for all SAMPL4 statistics. For all metrics the gains are significant, moving the FEP/ML prediction to 1st or 2nd rank as judged by MUE, r or Kendall tau metrics, and from 28th to 4th position as judged by RMSE. Many of the top-performing methods have very similar performance within statistical uncertainties, so care must be taken not to overinterpet changes in rankings. Nevertheless it is clear that the ML-derived correction terms improve the accuracy of the FEP methodology.

ML performed broadly similarly to FEP, but the uncertainty of the metrics is again remarkably large. This indicates that there is significant variability in the predicted free energies of hydration of the same compound by the ensemble of ML models. By contrast the FEP/ML predictions are of similar precision to the FEP predictions as the uncertainties in the offset terms is comparable or smaller to the uncertainties in the alchemical estimates.

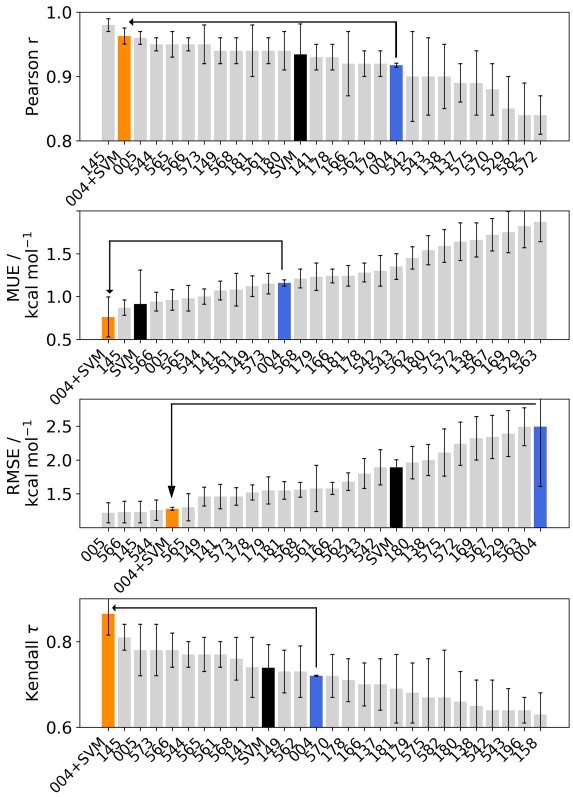

Figure 3: *SAMPL4 statistical performances of top-ranked entries with inserted pure ML and FEP/ML predictions as depicted in the original challenge.* Entry 004 (standalone FEP) is shown in blue. The FEP/ML model is shown in orange, and rank gains between standalone FEP and FEP/ML are depicted as black arrows. Pure ML models (ML) are shown as black bars. Error bars show model uncertainties as depicted in the SAMPL4 overview publication.[52]

Influence of training set size on accuracy of correction terms

We also evaluated the impact of training set size on the accuracy of the correction terms (figure 4). Hyperparameter configurations were taken from top performers in the training phase of this study (see table S1), and increasingly large, randomly sampled subsections of FreeSolv (exluding the test set) were used were used as training sets. For simplicity only SVM (trained using MolPropsAPFP) results are shown as this model consistently outperformed all others.

It was observed that with training sets of increasing size the cost function (in this case, MUE of FEP/ML prediction on SAMPL4 in kcal·mol^{-1}) decreases monotonically. FEP/ML models appear to outperform standalone FEP after being trained on ca. 20 compounds in FreeSolv (figure 4A), and converge with training sets of ca. 400 compounds. Strikingly, standalone ML models require much larger training sets of ca. 450 compounds to outperform standalone FEP. In both cases the gradual decrease in uncertainty with increase in training set size is due to higher overlap in training sets composition between replicates as the full training set size (n=595) is approached. Whereas the FEP/ML model seems to converge at ca. 400 compounds, the ML model does not appear to have converged and could likely benefit from a larger training set.

To put these results in perspective in the context of SAMPL4, the changes in ranks of the FEP/ML entry was plotted as a function of training set size (figure 4B). FEP/ML models outperform standalone FEP for all statistical measures, although some variability is observed. Whereas MUE and Kendall τ already show clear improvements from small training set sizes (ca. 100 and 50, resp.), Pearson r and RMSE appear to require models trained on a larger number of compounds to reach placement in the top five ranks of the SAMPL4 challenge (250 and 500, resp.).

A top-ranked result by Pearson r is not achieved even with a full training set of 595 compounds. This is also apparent in figure 3, where entry 145 is shown to outperform the FEP/ML model. This entry consists in a quantum-mechanical-based method with implicit

solvent and applies an empirical correction term to alcohol, ether, ester, amines and aromatic nitrogen groups which were derived from experimental data.[58] It is difficult to compare correction terms in this case because these corrections are generated from experimental measures versus Poisson-Boltzmann-based free energy calculations.

Although FEP/ML hybridisation does not appear to benefit RMSE scores in figure 3, the RMSE ranking for FEP/ML models appear to approach first place in the SAMPL4 challenge when trained on the full training set (595 compounds). The working model in figure 3 is trained using a cross-validation approach which effectively limits training set sizes to $0.8 * 595 = 476$ compounds which suggests that when generating a definitive ML correction term it would be preferable to use all 595 compounds as a training set.

The offsets are transferable to a number of related SAMPL4 submissions

The transferability of the ML-derived offsets to related simulation protocols was also assessed to evaluate the general applicability of the methodology. Figure 5 summarises changes in metric ranks for all complete submissions that featured an FEP methodology (n=19). Overall the offsets improved/maintained/worsen the rankings of 12/5/2 submissions for Pearson r; 10/3/6 submissions for MUE and RMSE; 9/6/4 submissions for Kendall Tau. Importantly with one exception (see below) the offsets do not worsen the ranks of the top-performing submissions.

As expected, SAMPL4 submission 004 is among the entries that benefit the most from the correction terms. Several entries that used a similar forcefield (GAFF and AM1-BCC charges, gromacs simulation engine) but a different simulation engine or different free energy estimation protocols (e.g. 137, 168, 544, 575) also show improvements in metrics. This is reasonable as it has been shown that, when properly implemented, hydration free energies computed with the same forcefield by different simulation engines will broadly agree to within $0.2 kcal \cdot \mathrm{mol}^{-1}$.[10]

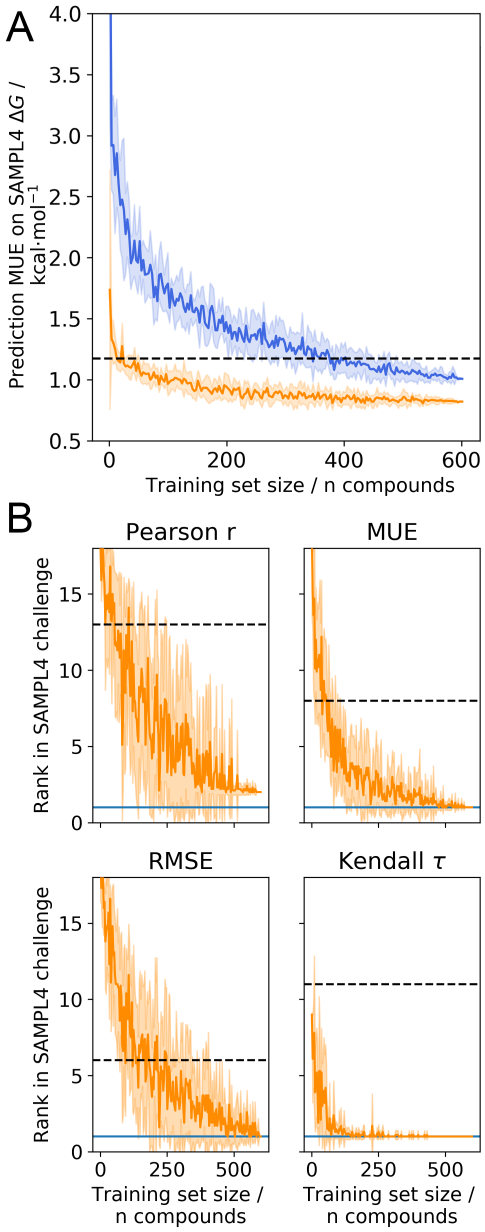

Figure 4: *Effect of increasing training set size on machine-learned correction models.* Results depicted are produced by support vector machines trained using MolPropsAPFP and MolProps for FEP/ML and pure ML models, respectively. **A:** FEP/ML model mean unsigned errors in the SAMPL4 challenge are shown with increasingly large (randomly sampled) subsets of the FreeSolv database as training sets with uncertainties across replicates (n=10) shown as lighter-shaded regions. Orange and blue lines are FEP/ML (FEP+ML, trained on ΔG_{offset}) and pure ML (trained on ΔG) predictors, respectively. Horizontal dashed line indicates the standalone FEP MUE of the FreeSolvSAMPL4 set in the SAMPL4 challenge. **B:** results for the same experiment as *A* but with ranking position of the FEP/ML model in the SAMPL4 challenge on the y axis per statistical measure. Horizontal dashed lines indicates the standalone FEP statistical measures of the FreeSolvSAMPL4 set in the SAMPL4 challenge and solid blue lines indicate first place in the challenge (i.e. $y = 1$).

The charge model used significantly influences the transferability of the offsets. Submission 542, 543, 545 only differ from submission 544 in the charge model used (RESP/HF-631G*, RESP/MP2/aug-cc-pVDZ/PCM, vCHARGE, AM1-BCC respectively). The offsets worsen the accuracy of the RESP methods but improve slightly the vCHARGE results. Other RESP-based submissions (166, 167, 169) see marginal changes in ranks. Submissions based on OPLS forcefields (562, 563, 564) benefit somewhat from the offsets, but not a GROMOS (529) or an AMOEBA (582) submission. This may be explained by the higher correlation of the AM1-BCC/GAFF hydration free energies with the OPLS hydration free energies (Pearson r 0.95, mean absolute deviation $1.1 kcal \cdot \mathrm{mol}^{-1}$) than the GROMOS hydration free energies (Pearson r 0.84, MUE 1.9 kcal $\cdot \mathrm{mol}^{-1}$) or AMOEBA hydration free energies (Pearson r 0.86, MUE 3.5 kcal$\cdot \mathrm{mol}^{-1}$).

A number of submissions made use of empirical correction terms that account for known deficiencies of the GAFF force field. For instance submission 005 corrects the tendency of the GAFF forcefield to underhydrate hydroxyls.[52,59] This source of error has been picked up by the ML-models as evidenced by the large offsets for mannitol (Figure 1B). Consequently applications of the offsets to submission 005 overshoots the hydration free energy of this compound, which contributes to a noticeable loss of ranks in MUE/RMSD. Submission 006 also includes an additional polarisation correction term that partially cancelled the effect of the hydroxyl correction term. The offsets restore partially the correction, leading to improved rankings. A similar behavior is observed with submission 138 that used QM derived corrections to improve GAFF hydration free energies reported in submission 137, leading to redundancy with the ML-derived offsets.

Conclusions

This work has demonstrated that it is possible to combine 'physics-driven' FEP methods with 'data-driven' machine learning methods to predict absolute hydration free energies of

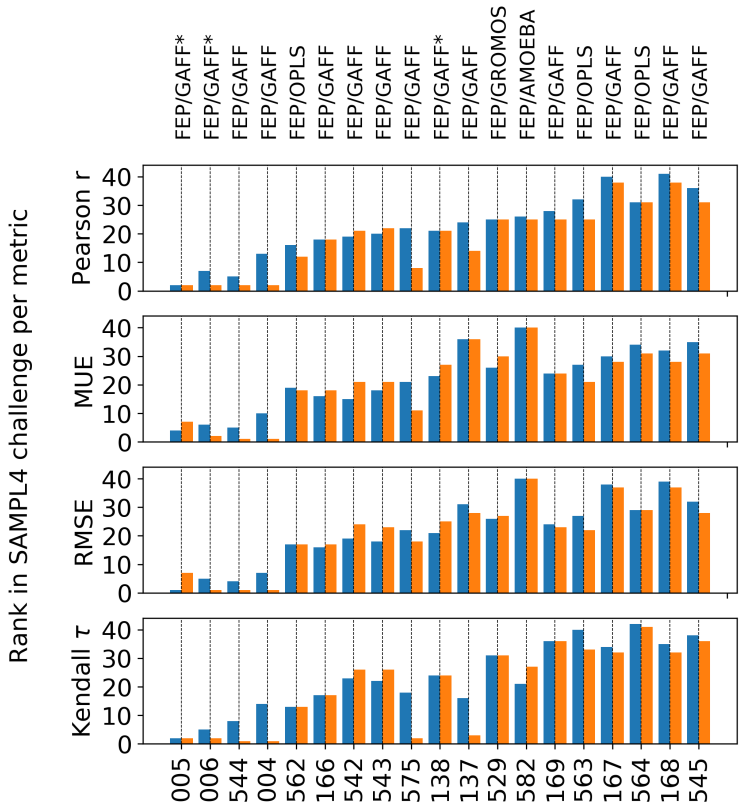

Figure 5: *Changes in ranks of SAMPL4 submissions after application of offsets to predicted hydration free energies.* Depicted are SAMPL4 FEP entries before (blue) and after (orange) hybridisation with the SVM-MolPropsAPFP correction term. The version of this plot with non-FEP entries can be found in figure S7. Entries were sorted by total ranks gained in ascending order. The FreeSolvSAMPL4 set corresponds to entry 004.

small molecules. The chief advantage over FEP is that improvements in the accuracy of the predictions are achieved without having to embark in cumbersome forcefield parameterization efforts. When compared with ML, the FEP/ML approach outperforms FEP with a much smaller training set size. This is significant as it indicates that for a new dataset it is possible to make predictions without any available experimental data initially, and switch to an FEP/ML approach once a sufficient number of data points have been experimentally determined. This advantage stems from the fact that in the FEP/ML approach the ML models only need to learn to correct errors in the FEP calculations, whereas in a pure ML approach the models must learn the physics of hydration. Another advantage of FEP/ML is that the hydration free energies of individual compounds are predicted with precision similar to that of the FEP calculations, whereas ML-based predictions by ensemble of identical models show more significant variability. In a retrospective analysis of all SAMPL4 submissions, the accuracy gains obtained in FEP/ML are sufficient to propel a mid-ranked FEP protocol among the top-ranked submissions. Further, the accuracy improvements are not limited to a single simulation protocol, and a number of related FEP approaches benefit from the correction terms. This likely stems from the fact that the hydration free energies predicted by a number of forcefields and software show correlations in their outliers.[10,60] However the performance of the correction terms is expected to decrease the more the simulation protocol diverges from that used to generate the training set.

There would be of course no need for such correction terms if more accurate forcefields were available. Thus beyond empirically correcting forcefield errors, the ML correction terms are useful to flag at essentially no computing cost molecules for which predictions are likely to deviate significantly from experimental data. This should be useful to help focus time-consuming forcefield parameterization efforts, or as part of automated workflows to decide whether to embark in bespoke forcefield parameterization for a given compound. Finally, the methodology presented here could be applied to other scenarios where FEP is used extensively, for instance relative or absolute protein-ligand binding free energy calculations. This

will likely require further methodological developments to handle non negligible statistical sampling errors in the FEP results; as well as learning of a diverse set of physical interactions present in the more heterogeneous environment found in protein binding sites. Nevertheless the current growth in size and diversity of protein-ligand datasets with associated FEP data should render FEP/ML an increasingly appealing option to improve the effectiveness of FEP methods in drug discovery.[12,14,61]

Supporting Information Available

Additional figures and tables. Jupyter notebooks to generate the models and figures reported in this study. Scripts and inputs are also available at `https://github.com/michellab/hybrid_FEP-ML`.

Conflict of Interest

JM is a current member of the Scientific Advisory Board of Cresset.

References

(1) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.

(2) Simone, A. D.; Georgiou, C.; Ioannidis, H.; Gupta, A. A.; Juárez-Jiménez, J.; Doughty-Shenton, D.; Blackburn, E. A.; Wear, M. A.; Richards, J. P.; Barlow, P. N.; Carragher, N.; Walkinshaw, M. D.; Hulme, A. N.; Michel, J. A computationally designed binding mode flip leads to a novel class of potent tri-vector cyclophilin inhibitors. *Chem. Sci.* **2019**, *10*, 542–547.

(3) Kuhn, B.; Tichý, M.; Wang, L.; Robinson, S.; Martin, R. E.; Kuglstatter, A.; Benz, J.; Giroud, M.; Schirmeister, T.; Abel, R.; Diederich, F.; Hert, J. Prospective Evaluation of Free Energy Calculations for the Prioritization of Cathepsin L Inhibitors. *J. Med. Chem.* **2017**, *60*, 2485–2497.

(4) Georgiou, C.; McNae, I.; Wear, M.; Ioannidis, H.; Michel, J.; Walkinshaw, M. Pushing the Limits of Detection of Weak Binding Using Fragment-Based Drug Discovery: Identification of New Cyclophilin Binders. *J. Mol. Biol.* **2017**, *429*, 2556–2570.

(5) Michel, J. Current and emerging opportunities for molecular simulations in structure-based drug design. *Phys. Chem. Chem. Phys.* **2014**, *16*, 4465–4477.

(6) Mishra, S. K.; Calabró, G.; Loeffler, H. H.; Michel, J.; Koča, J. Evaluation of Selected Classical Force Fields for Alchemical Binding Free Energy Calculations of Protein-Carbohydrate Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3333–3345.

(7) Chen, I.-J.; Foloppe, N. Is conformational sampling of drug-like molecules a solved problem? *Drug Dev. Res.* **2011**, *72*, 85–94.

(8) Rocklin, G. J.; Mobley, D. L.; Dill, K. A. Calculating the Sensitivity and Robustness of Binding Free Energy Calculations to Force Field Parameters. *J. Chem. Theory Comput.* **2013**, *9*, 3072–3083.

(9) Calabrò, G.; Woods, C. J.; Powlesland, F.; Mey, A. S. J. S.; Mulholland, A. J.; Michel, J. Elucidation of Nonadditive Effects in Protein–Ligand Binding Energies: Thrombin as a Case Study. *J. Phys. Chem. B* **2016**, *120*, 5340–5350.

(10) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.

(11) Assessment of Binding Affinity via Alchemical Free Energy Calculations. *J. Chem. Inf. Model.* **2020**, in press.

(12) Wang, L. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(13) Song, L. F.; Lee, T.-S.; Zhu, C.; York, D. M.; Merz, K. M. Using AMBER18 for Relative Free Energy Calculations. *J. Chem. Inf. Model.* **2019**, *59*, 3128–3135.

(14) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; Vlijmen, H. v.; Tresadern, G.; Groot, B. L. d. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **2020**, *11*, 1140–1152.

(15) Rizzi, A. et al. The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations. *J. Comput.-Aided Mol. Des.* **2020**,

(16) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 host–guest challenge: Are we doing better? *J. Comput.-Aided Mol. Des.* **2017**, *31*, 1–19.

(17) Parks, C. D.; Gaieb, Z.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Jansen, J. M.; McGaughey, G.; Lewis, R. A.; Bembenek, S. D.; Ameriks, M. K.; Mirzadegan, T.; Burley, S. K.; Amaro, R. E.; Gilson, M. K. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 99–119.

(18) Mey, A. S. J. S.; Jiménez, J. J.; Michel, J. Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 199–210.

(19) Granadino-Roldán, J. M.; Mey, A. S. J. S.; González, J. J. P.; Bosisio, S.; Rubio-Martinez, J.; Michel, J. Effect of set up protocols on the accuracy of alchemical free energy calculation over a set of ACK1 inhibitors. *PLOS ONE* **2019**, *14*, e0213217.

(20) Papadourakis, M.; Bosisio, S.; Michel, J. Blinded predictions of standard binding free energies: lessons learned from the SAMPL6 challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1047–1058.

(21) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.

(22) Martins, S. A.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Prediction of Solvation Free Energies with Thermodynamic Integration Using the General Amber Force Field. *J. Chem. Theory Comput.* **2014**, *10*, 3570–3577.

(23) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 4533–4537.

(24) Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 927–944.

(25) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528–552.

(26) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.

(27) Slochower, D. R.; Henriksen, N. M.; Wang, L.-P.; Chodera, J. D.; Mobley, D. L.; Gilson, M. K. Binding Thermodynamics of Host–Guest Systems with SMIRNOFF99Frosst 1.0.5 from the Open Force Field Initiative. *J. Chem. Theory Comput.* **2019**, *15*, 6225–6242.

(28) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076–6092.

(29) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* **2019**, *48*, 371–394.

(30) Beierlein, F. R.; Michel, J.; Essex, J. W. A Simple QM/MM Approach for Capturing Polarization Effects in ProteinLigand Binding Free Energy Calculations. *J. Phys. Chem. B* **2011**, *115*, 4911–4926.

(31) König, G.; Pickard, F. C.; Mei, Y.; Brooks, B. R. Predicting hydration free energies with a hybrid QM/MM approach: an evaluation of implicit and explicit solvation models in SAMPL4. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 245–257.

(32) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(33) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(34) Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* **2017**, *7*, 46710.

(35) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 405–424.

(36) Lim, H.; Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* **2019**, *10*, 8306–8315.

(37) Lim, H.; Jung, Y. MLSolv-A: A Novel Machine Learning-Based Prediction of Solvation Free Energies from Pairwise Atomistic Interactions. *arXiv:2005.06182 [cond-mat, physics:physics, stat]* **2020**,

(38) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726–741.

(39) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1338–1346.

(40) Ramsundar, B.; Eastman, P.; Patrick Walters,; Pande, V.; Karl Leswing,; Zhenqin Wu, *Deep Learning for the Life Sciences*; OReilly: Sebastopol, CA, US, 2019.

(41) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.

(42) Jiménez-Luna, J.; Pérez-Benito, L.; Martínez-Rosell, G.; Sciabola, S.; Torella, R.; Tresadern, G.; Fabritiis, G. D. DeltaDelta neural networks for lead optimization of small molecule potency. *Chem. Sci.* **2019**, *10*, 10911–10918.

(43) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(44) Bosisio, S.; Mey, A. S. J. S.; Michel, J. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 61–70.

(45) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(46) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated

hydration free energies, with input files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711–720.

(47) Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62*, 1559–1569.

(48) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(49) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

(50) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(51) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(52) Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. Blind prediction of solvation free energies from the SAMPL4 challenge. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 135–150.

(53) Landrum, G. RDKit: Open-source cheminformatics. 2020; `https://github.com/rdkit/rdkit`.

(54) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, 4.

(55) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(56) Head, T. et al. scikit-optimize/scikit-optimize: v0.5.2. 2018; `https://zenodo.org/record/1207017#.XNWN045KhaQ`.

(57) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(58) Sandberg, L. Predicting hydration free energies with chemical accuracy: the SAMPL4 challenge. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 211–219.

(59) Fennell, C. J.; Wymer, K. L.; Mobley, D. L. A fixed-charge model for alcohol polarization in the condensed phase, and its role in small molecule hydration. *J. Phys. Chem.. B* **2014**, *118*, 6438–6446.

(60) Bosisio, S.; Mey, A. S. J. S.; Michel, J. Blinded predictions of distribution coefficients in the SAMPL5 challenge. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 1101–1114.

(61) Schindler, C. et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *ChemRxiv* **2020**,

Graphical TOC